Combinatorial Genetic Regulatory Network Analysis Tools for High Throughput Transcriptomic Data

Elissa J. Chesler[1] and Michael A. Langston[2]

[1] Life Sciences Division, Oak Ridge National Laboratory,
P.O. Box 2008, Oak Ridge, TN 37831-6124, USA
[2] Department of Computer Science, University of Tennessee,
Knoxville, TN 37996–3450, USA

Abstract: A series of genome-scale algorithms and high-performance implementations is described and shown to be useful in the genetic analysis of gene transcription. With them it is possible to address common questions such as: “are the sets of genes co-expressed under one type of conditions the same as those sets co-expressed under another?” A new noise-adaptive graph algorithm, dubbed “paraclique,” is introduced and analyzed for use in biological hypotheses testing. A notion of vertex coverage is also devised, based on vertex-disjoint paths within correlation graphs, and used to determine the identity, proportion and number of transcripts connected to individual phenotypes and quantitative trait loci (QTL) regulatory models. A major goal is to identify which, among a set of candidate genes, are the most likely regulators of trait variation. These methods are applied in an effort to identify multiple-QTL regulatory models for large groups of genetically co-expressed genes, and to extrapolate the consequences of this genetic variation on phenotypes observed across levels of biological scale through the evaluation of vertex coverage. This approach is furthermore applied to definitions of homology-based gene sets, and the incorporation of categorical data such as known gene pathways. In all these tasks discrete mathematics and combinatorial algorithms form organizing principles upon which methods and implementations are based.

Keywords: Microarray Analysis, Putative Co-Regulation, Quantitative Trait Loci, Regulatory Models

1 Introduction

We describe ongoing research efforts aimed at developing, implementing and validating graph theoretical approaches and high-performance computing implementations for the systems genetic analysis of high throughput molecular phenotypes in relation to higher order systems-level traits. Present approaches to these problems typically deal only with individual genes or small sets of genes and a small handful of systems phenotypes. Early analytic approaches relied primarily on simplifying assumptions that only a single locus is involved in gene regulation [12, 18, 36]. This is despite widespread acknowledgment that gene expression is a complex phenotype regulated by multiple genetic and environmental factors. It has been simply a limitation of the commonly employed analytic tools, which only evaluate one locus at a time. Recent studies have systematically examined two-locus interactions [11], but we have observed regulation by larger combinations of loci. To search the model space for the best mufti-locus modeling, arbitrary filtering

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **2006**	2. REPORT TYPE	3. DATES COVERED **00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE **Combinatorial Genetic Regulatory Network Analysis Tools for High Throughput Transcriptomic Data**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Tennessee,Department of Computer Science,Knoxville,TN,37996-3450**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**17**	

of the data is often required to reduce the problem size to a manageable size. Instead, we represent the entire data set in a graph theoretical context, and employ a unique top-down approach. We transform the expression genetic covariance matrix into a simple, undirected, unweighted graph, and extract pure groups of highly inter-correlated expression traits. These groups represent a much smaller set of traits that are then subject to mapping analysis. We examine the connectivity among these sets and analyze the molecular, biochemical and genetic regulatory commonality of connected genes using novel and existing bioinformatics tools. We also develop data-driven hypotheses to explain the mechanisms of genetic perturbations and variation as a means of defining global consequences of individual differences on tissue structure and function.

Much of our work is motivated by prior studies of brain gene expression and mRNA abundance levels. These are complex phenotypes regulated by multiple genetic and environmental factors [21]. We have previously performed genome-wide mapping [3] of the loci that modulate brain gene expression assayed on microarrays in a genetic reference population, and explored correlations of expression and complex phenotypes [20]. These and other transcriptome mapping studies have revealed several major "trans-bands," that is, loci that regulate large numbers of transcripts encoded across the genome [33]. For example, we have recently identified at least seven loci that act in combination to regulate large numbers of transcripts encoding synaptic proteins and transcription or translation machinery [18]. Previous studies of genetic analysis of gene expression report trans-bands that reflect a high degree of covariance in the gene expression data [12, 16, 30, 36].

2 Quantitative Trait Loci and Regulatory Models

The sources of genetic covariation, used here to define edge weights, are the genetic polymorphisms that occur naturally among individuals. These differences, acting first on the molecular scale, exert their effects through time, space and tissue compartment to influence traits as diverse as morphology, physiology and behavior. Because the genetic variation and covariation that we seek to explain is continuous (or quantitative) in nature, they are referred to as quantitative traits, determined by multiple genes and environmental conditions. To identify the genes (here literally referring to the heritable source of variation, not strictly to a particular class of genome features), experimental crosses are performed to shuffle genotypes from two genomes through meiotic recombination. For example, mouse strains C57BL/6J (B) and DBA/2J (D) are crossed, generating an F1 population with one copy of the B allele and one copy of the D allele at every location throughout the genome. These mice are crossed again to create an F2 generation, each member of which has a unique complement of shuffled B and D genomes, that is, they possess either two B alleles, a B and a D allele, or two D alleles at each locus. Quantitative techniques have been developed to associate the vectors of polymorphic states (genotypes) at known locations throughout the genome with the vector of phenotypes [23, 34]. Statistically significant predictive genotype-phenotype

[3] Our transcriptomic data is publicly accessible in the WebQTL system www.webqtl.org. This tool allows systems genetic analysis of single genes or small sets of genes using a bottom-up approach.

relations define quantitative trait loci (QTLs). Because the marker is not typically the actual site of the polymorphism, interpolative methods have been developed to estimate the distance of the QTL from the marker and the strength of the association. Using multiple-regression and model-fitting methods, the true complexity of the phenotypic variation can be modeled through the consideration of multiple loci and environmental factors as predictors [13].

The typical experimental mapping population is bred once for each experiment, and the unique assortment of genotypes that result can never be retrieved. Mouse panels have been in use since the early 1980's, however, that can be used as a retrievable reference population. These populations, called recombinant inbred lines (henceforth Rill) consist of the progeny of an F2 cross that have been inbred for over twenty generations. The importance of these panels is that they allow population genetics methods to be applied to systems biology, by exploiting naturally occurring polymorphisms to determine the membership of biological networks that transcend time, space and tissue compartment. The largest of these sets contains eighty lines [35], and a large 1024 strain set is being bred at the Oak Ridge National Laboratory [22]. The genotypes are obtained at a very high precision and stored in public repositories such as www.genenetwork.org (encompassing WebQTL [19]) for analysis. Recently, genotypes have been identified at 15,000 loci using the Illumina SNP genotyping system (http://www.well.ox.ac.uk/mouse/INBREDS/RIL/index.shtml). Because the lines are inbred, attributes of the lines, whether they be genotypes, phenotypes, or high precision molecular trait data including microarray, can be aggregated indefinitely to form a single data matrix and analyzed using correlation techniques [20]. Phenotypic correlations can be partitioned into environmental and genetic sources of covariance. When the correlations are obtained on the phenotypic means based on subsampling of individuals within line, they can often be interpreted to be genetic correlations. This is particularly true if the trait data are obtained in independent sets of individuals. In transcription profiling, the multiple measures are observed in the same individuals, and it is possible, especially with low sample sizes, that these correlations are also largely environmentally driven.

The application of QTL mapping to microarrays, often termed "genetical genomics," was first reported in yeast [12], and has since been successfully performed in F2 [36] and RIL [16, 18] mouse populations. In recent work [18], we have detected the presence of trans-QTL bands (locations in the genome that regulate hundreds of distally located transcript abundances) that regulate co-expression in the central nervous system. The trans-QTL bands are typically found by fitting single-locus models across the genome, and then, at locations throughout the genome, counting the number of transcripts for which the best fitting peak is found at that location. This bottom-up approach requires several assumptions that do not always hold, notably, that each transcript abundance is regulated by a single genetic locus. We have taken a top-down approach to this problem. We use graph algorithms first to decompose the genetic correlation matrix so as to identify sets of putatively co-expressed phenotypes, and then to determine the best multiple locus models to determine their regulation. Each of the sets of phenotypes that we have identified is regulated by a combination of the trans-QTL bands. This top-down approach yields tremendous advantages due to the enormous computational demands that would be incurred if searching the entire model space.

Fundamental to this approach is the use of what we call “paracliques.” Informally, a paraclique is an extremely densely-connected subgraph, but one that may be missing a small number of edges and thus is not, strictly speaking, a clique. In the present application, this corresponds to a very highly intercorrelated group of genetically co-regulated genes whose transcript expression levels, as reflected in real and surely somewhat dirty microarray data, show highly significant but not necessarily perfect pair-wise correlations. By harnessing the computational power of tools such as fixed-parameter tractability, and then isolating paracliques, we are able to identify considerably denser subgraphs than are typically produced with traditional clustering algorithms. We have therefore reduced the immense genetic correlation matrix to a select set of intercorrelated modules, and have greatly simplified the discovery of functional significance and identity of genetic polymorphisms that underlie gene expression variation.

The RILs have been screened on over 1500 diverse phenotypes, and at least ten gene expression microarray profiling studies are completed or in progress in the RILs, which include several mouse, rat and plant species. The coverage of each paraclique (to be introduced in the sequel) by genotypes is determined. In this case, p-values are used as edge weights, due to the diverse sample sizes and correlation metrics applied to the data. The result is a graph of gene-to-phenotype relations. It consists of phenotype to expression relations, co-expression to regulatory models, and models to independent loci. See Figure 1.

3 Clique, Putative Co-Regulation and Fixed-Parameter Tractability

We adopt graph theoretical approaches to analyze the huge correlation matrix that results from a microarray experiment that records expression levels over thousands of probes conducted over many conditions. The matrix is transformed into a complete graph, in which each gene is represented by a vertex, and in which each edge is weighted by the correlation coefficient of its endpoints. A suitable threshold, t, is then chosen. We are currently employing a variety of techniques to make this selection [32]. These include the use of functional similarity, ontological enrichment, known gene product interactions, and even methods based on spectral graph theory. A high-pass filter is applied to eliminate any edge whose weight is less than t. At this point, the weight of any remaining edge is ignored. This procedure produces a simple, undirected, unweighted graph for subsequent study. Note that all genes remain in the analysis. It is only the weakest correlations that are discarded.

A clique [10] in this new graph denotes a set of genes with the interesting property that every pair of its elements is highly correlated. This is widely interpreted as suggestive of putative co-regulation over the conditions in which the experiment was performed. Clique finding can be viewed as an especially stringent graph-theoretical form of clustering for gene co-expression data. Although clique presents an exceedingly difficult computational problem, its advantages are many. It is particularly noteworthy that a vertex can reside in more than one clique, just as a gene can participate in more than

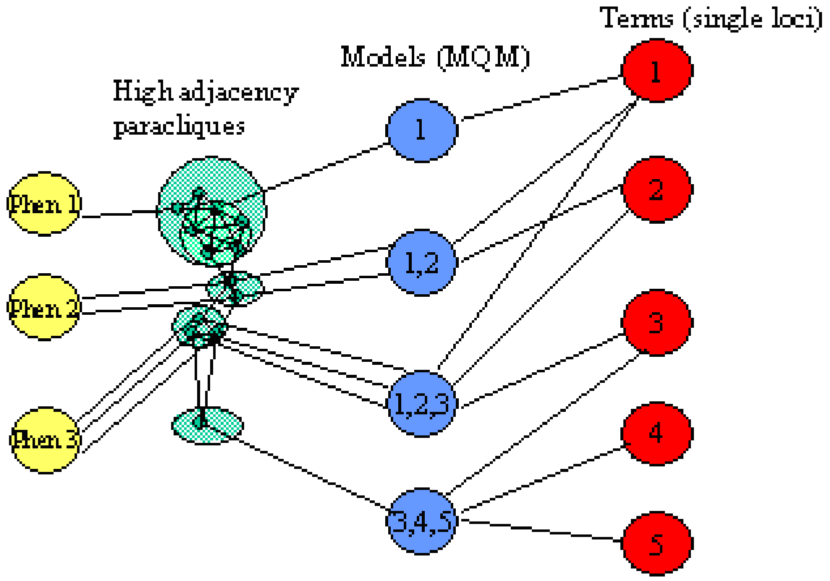

Fig. 1. Systems genetics can be viewed as a graph theoretical problem. With gene expression analysis (green), groups of co-expressed genes aka cliques are assembled into larger, disjoint paracliques. These can then be examined for coverage to phenotypes (yellow) and QTL models (blue) made of multiple single regulatory loci (red).

one regulatory network. A huge variety of other clustering strategies are known that attempt to organize multivariate data into groups with approximately similar expression patterns [7–9, 27–29, 37]. Like ours, most methods build upon a correlation measure between expression levels to calculate a distance metric of similarity (or dissimilarity) of expression between each gene pair. There are several important limitations, however, to the vast majority of clustering algorithms that lie in contrast to the realities of biology. One such limitation is that the clusters these methods produce are disjoint, requiring that a gene be assigned to only one cluster. While this simplifies the amount of data to be evaluated, it places an artificial limitation on the biology under study because many genes play important roles in multiple but distinct pathways [14]. There are recent clustering techniques, for example those employing factor analysis [3], that do not require exclusive cluster membership for single genes. Unfortunately, these tend to produce biologically uninterpretable factors without the incorporation of prior biological information [26]. Another important limitation is that most of the measures of similarity used by current clustering algorithms do not permit the recognition of negative correlations, which are common and often equally meaningful from a biological perspective.

Ours is in fact a very general approach. Correlations between mRNA abundances and other molecular phenotype levels can be used as well to describe an edge-weighted graph, in which each vertex represents a measured attribute (e.g., gene expression, behavior or genotype), and in which an edge between two attributes is weighted with the appropriate correlation coefficient. As with microarray data, this graph can be enormous. It is built from the entire trait x trait correlation matrix. A variety of graph algorithms can now be applied, in particular those designed to extract densely-connected subgraphs. By definition the most densely connected subgraphs of all are of course again cliques. A clique in this context may help identify an important biological module (e.g., a subunit of a protein complex). Alternately, it may point to correlation only when a strong driving biological force causes its elements to co-vary. Clique-centric tools provide us with powerful techniques for the study of highly interconnected groups of traits.

We typically seek all maximal cliques[4], but it can be folly to try to compute them without first knowing a bound on their size. To do this, we solve maximum clique with the aid of fixed-parameter tractability (FPT).

> A problem is FPT if it has an algorithm that runs in $O(f(k)n^c)$ time, where n is the problem size, k is the input parameter, and c is a constant independent of both n and k.

Clique is not FPT, however, unless the W hierarchy[5] collapses [24]. Thus, we focus instead on clique's complementary dual, the vertex cover problem, and on G', the complement of G. (G' has the same vertex set as G, but edges present in G are absent in G' and vice versa.) Both clique and vertex cover are $\mathcal{NP}$-complete. Unlike clique, however, vertex cover is FPT. The relevant observation here is this: a vertex cover of size k in G' turns out to be exactly the complement of a clique of size $n-k$ in G. We therefore search for a minimum vertex cover in G', thereby finding the desired maximum clique in G. Currently, the fastest known vertex cover algorithm runs in $O(1.2759^k k^{1.5} + kn)$ time [17]. The requisite exponential growth (modulo $\mathcal{P} \neq \mathcal{NP}$) is thus reduced to a mere additive term, making it realistic now to consider the search for cliques of immense sizes. Some of our recent progress on FPT and maximum clique is featured in [1, 2]. Our latest work on high performance solutions to maximal clique enumeration can be found in [39].

4 Noise, Overlap and the Paraclique Method

Clique is an ideal cluster definition, the gold standard. Every vertex (transcript) in a clique must be highly correlated with all others in that clique by definition. Microarray data, on the other hand, at least as generated under current technologies, are inherently noisy. It seems highly unlikely that noise alone could cause an entire clique's worth

[4] A maximal clique in a graph G is one that is locally optimal. That is, it is a complete subgraph with the property that no new vertex in G can be added to it. It should not be confused with the more common notion of maximum clique, which is a largest clique in G.

[5] The W hierarchy, whose lowest level is FPT, can be viewed as a fixed-parameter analog of the polynomial hierarchy, whose lowest level is $\mathcal{P}$. Such a collapse is widely viewed as an exceedingly unlikely event, roughly on a par with the likelihood of the collapse of the polynomial hierarchy.

of correlation coefficients to be excessively high, whereas common clustering methods including KNN and K-Means can allow transcripts into a group that are related due to noise. Thus, in contrast to many other popular approaches, clique does not seem to be plagued with false positives. On the other hand, if even one coefficient is incorrectly found to be too low, then the clique is lost, equaling a false negative. Several edges may even be missing from the largest subgraphs. The result is that a typical clique analysis may yield exceedingly large numbers of modest-sized, highly-overlapping cliques [31]. To aggregate these data, we want to solve something akin to dense-k-subgraph [25], which is $\mathcal{NP}$-complete even on graphs of maximum degree three.

To accomplish this we have developed a novel algorithmic approach that we term paraclique. Roughly speaking, a paraclique is a clique augmented with vertices in a highly controlled manner to maintain density. In what follows we describe the paraclique algorithm in pidgin Algol. It uses what we term a glom factor to latch onto new vertices, and an optional threshold to check the original weights of edges discarded by the high pass filter.

```
paraclique (graph $G$, glom factor $g$, threshold $t$)
set $P$ to $C$, some maximum clique in $G$
set $P'$ to $\emptyset$
    while $P \neq P'$ do
    set $P'$ to $P$
    for every $v \in V - P$ do
        if $v$ is adjacent to at least $g$ members of $P'$
            then if the weight of each edge connecting $v$ to $P'$ before filtering is at least $t$
                then set $P$ to $P \cup v$
        end for
    end while
return $P$
```

Although our interest is focused on real and not synthetic data, it is not at all difficult to prove the following.

Theorem. For any graph G, with g set to $|C| - 1$ the edge density of P as computed by the paraclique algorithm is at least 50% as long as $|P| \leq 2|C|$.

Paraclique can be iterated as long as needed, excising from G the current value of P at each pass. Empirical testing on real microarray data has been revealing. If one merely augments a clique with one- and two-neighborhoods, the edge density rapidly falls to the $10 - 20\%$ range. But with paraclique, if we set g to $|C| - 1$, then edge density tends to remain above 90% (even with t set to 0). All this is accomplished while the sizes of the paracliques generated are roughly twice the sizes of the respective cliques from which each was constructed.

5 Sample Results

We have applied our methods to mouse brain gene expression data collected by Robert W. Williams and colleagues and described in [18]. In the genetic analysis of this data, several major trans-regulatory QTLs were identified using a bottom up approach. Our first application of clique-centric analysis [4] to this data was performed on MAS5.0 normalized brain gene expression data using Spearman's rank correlations. We began with a threshold setting of 0.50 and, using our FPT-based algorithms, consumed roughly a week of CPU time on a 32-processor cluster to determine that the maximum clique size was 369. Analyzing cliques of this size was deemed to be beyond the current capability of effective biological verification methods. We therefore iterated over a variety of thresholds until we settled on a relatively high $|r| \geq 0.85$ value to filter the edge-weighted graph. This analysis revealed 5227 maximal cliques with a maximum clique size of 17. See Figure 2.

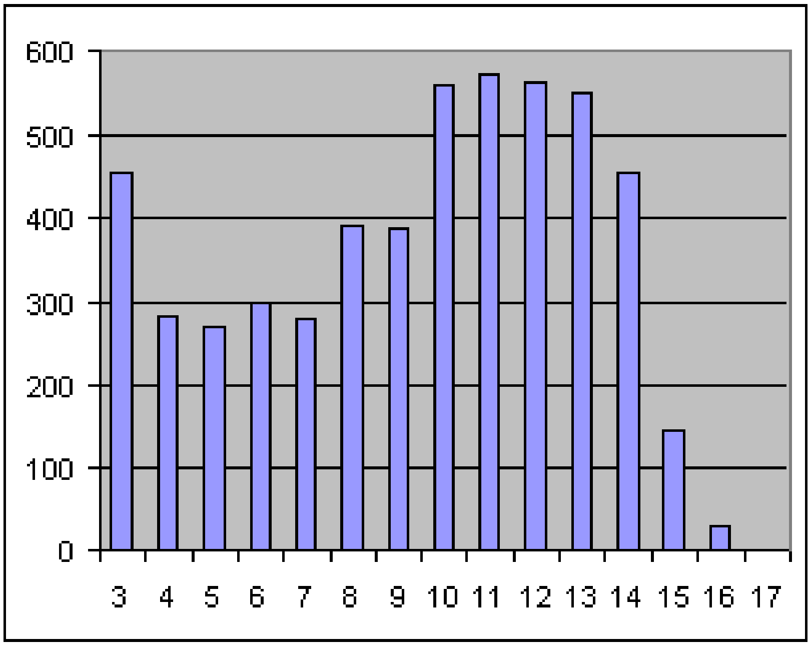

Fig. 2. Clique size distribution in MAS5.0 normalized brain.

The gene whose transcript is represented by far in the greatest number of maximal cliques is *Lin7c* (also knows as *Veli3*). This gene forms a complex with clique member *Cask* and *Mint1* to participate in the coupling of cell adhesion models to synaptic vesicle cycling [15]. This complex plays a role in the cycling of numerous neurotransmitter receptors, including *Htr2c* [6]. Other *Lin7c* clique members include *Gs2na*, which is a synapse and soma localized gene also involved in protein-protein interactions with consequences on locomotor behavior [5]. For this dataset it is particularly promising that clique-centric analysis has identified groups of genes encoding proteins that physically interact at the synapse. These cliques are highly overlapping. Simple array noise can often be responsible for separating a group of inter-correlated transcripts into a huge number of highly similar cliques, each varying due to the presence of a small number of missing edges. The annotation and interpretation of such a result is quite challenging. This challenge is even more pronounced with RMA normalizations, because of an overall increase in correlation coefficient values and with it an increase in graph density.

In subsequent analyses, we again studied the aforementioned brain gene expression data. We kept the threshold at 0.85, but used the RMA normalization package due to its greater precision. Correlation graph density is greatly affected by RMA. The maximum clique size increased from 17 to 280; the number of maximal cliques increased from 5227 to a value in excess of 9.5 million (where we stopped counting). Dimensional reduction using paraclique produced highly-purified gene sets, while maintaining densely-connected subgraphs and consolidating the overwhelming number of overlapping cliques. Millions of maximal cliques were reduced to a mere 31 paracliques, ranging in size from 12 to 466, with each paraclique having an edge density in excess of 95%. Thus, as designed, the paraclique algorithm is an attempt to correct for noise. It extracts dense, disjoint subgraphs. By definition, the majority of the interconnections among all transcripts must be present. This does not preclude the presence of edges between paracliques. We have observed that these dense subgraphs are marked by interconnectivity at interfaces between only a few vertices. Thus there are some vertices with high-coverage of multiple paracliques. The corresponding genes are likely to be important players in the regulatory networks with interconnections to these transcripts.

6 Impact

Mapping the QTL regulators of paraclique expression reveals that trans-QTL bands do not act independently, but rather, they act in concert to regulate simultaneously over 1700 transcripts. The paraclique algorithm has given us an unprecedented and simultaneous view of all of transcriptome QTL data. Using a derivative of the cluster map display designed for WebQTL [19], it is possible to visualize the combinations of trans QTL bands that are responsible for regulation of the paracliques. Using paraclique, we were able to decompose the genetic co-expression matrix into groups of transcripts with shared regulatory architecture, and have demonstrated that the trans-bands act in concert, as opposed to singly to regulate transcription. This result could not be obtained using a bottom up approach, which presupposes single regulatory loci, although we did see indications of this structure by simply mapping many functionally related transcripts in parallel. See Figure 3.

Once these key loci are isolated, the challenge is to identify the actual pathways and genes that are involved in biological networks that are perturbed by the genetic polymorphism. A plethora of annotation tools aimed at understanding gene sets have emerged over the past few years. The vast majority of genes that are co-expressed in the

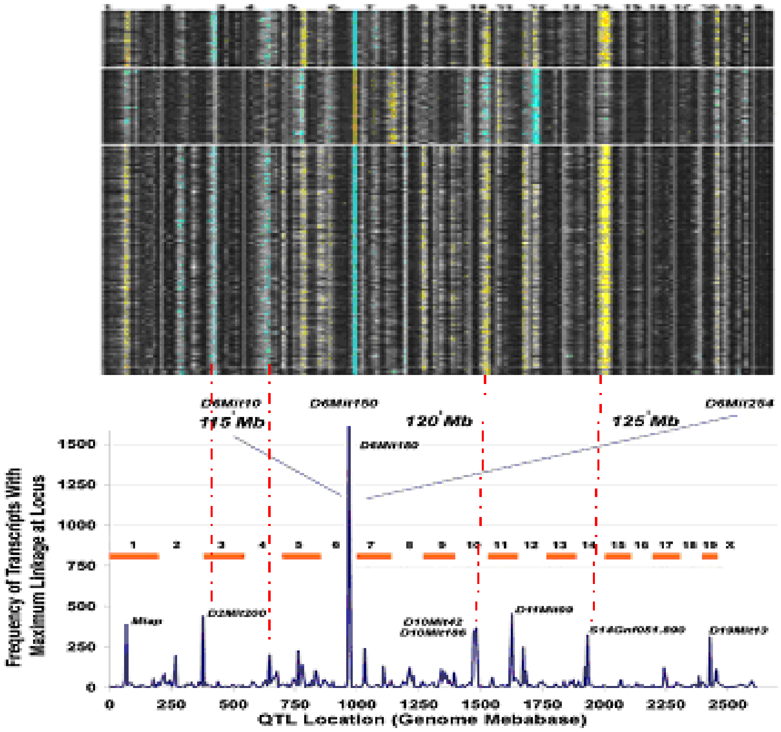

Fig. 3. Trans-QTL bands as shown in the lower panel are compared to the genetic regulation of paracliques as shown in the upper panel. The mouse genome is represented on the x-axis of both plots. The lower plot is courtesy of Nature Genetics [18]. In the upper plot, QTL members of three paracliques are plotted in parallel. Warm colors represent locations where the DBA/2J alleles decrease in expression levels. Cool colors represent locations where C57BL/6J alleles increase in expression levels. Each paraclique is regulated by a unique combination of trans-QTL loci.

paraclique graph are related to the neuronal synapse, and in particular the transport and translation of mRNAs at the dendrites. One of the compelling candidate genes for the regulatory locus at chromosome 1 is *Mtap2*, which is located in the QTL region, and is an expression correlate of transcripts in the region. See Figure 4.

7 Candidate Gene Selection

A QTL may contain tens to hundreds of genes, as was illustrated in Figure 3. The identification of candidate genes in a QTL region can be aided by an analysis of multiple converging evidences so as to rank these genes. Let us discuss just two of these evidences.

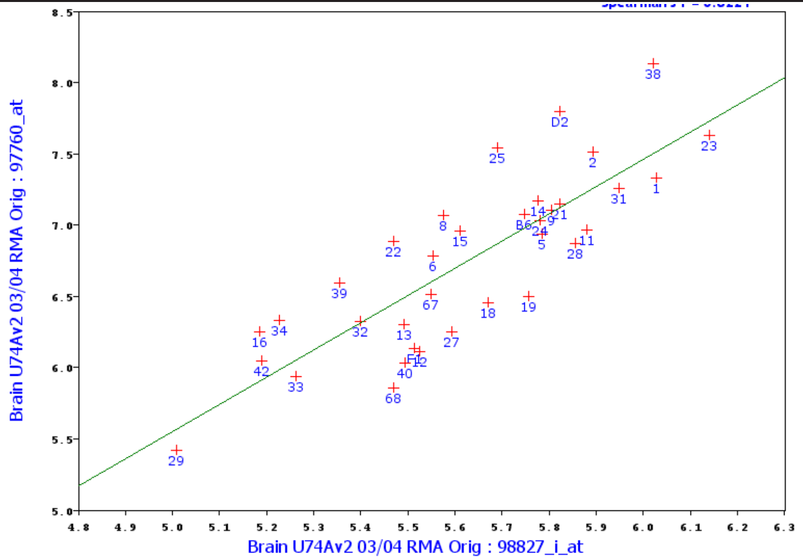

Fig. 4. An example of genetic correlation scatter plot revealing a high correlation between *Kif5a* and *Mtap2* abundance.

In the first approach, we integrate simple edge densities with genomic data to identify candidate regulators. We have observed that a small number of vertices within paracliques are highly adjacent to vertices in other paracliques. By examining the percentage of possible adjacent edges observed among paraclique members, we are able to identify the vertices at the interface of two or more paracliques. By further overlaying information regarding the genomic location of the transcripts represented by these vertices and comparing this position to the location of QTLs, we are able to identify paraclique members that contain the causative polymorphisms responsible for the covariance of paracliques. This approach is not exclusionary, nor is it fully inclusionary. This is because expression data is not available for some genes in the QTL interval. Moreover, in other cases, the polymorphism has functional effects on the gene but does not influence transcript abundance.

In the second approach, we borrow a page from evolutionary theory, which suggests that genes residing in one of many redundant paths are more mutable than those that are in exclusive pathways [38]. Suppose vertex a is a transcript abundance or higher order phenotype, and suppose genes b, c, d, e and f are the neighbors of a (determined by genetic correlation of transcript abundance to this phenotype in RI lines). We seek to find redundant pathways, aka vertex-disjoint paths, in a's gene neighborhood. See Figure 5.

Consider, for example, genes b and e. They may be adjacent via a path of length one, as shown in green. Alternately, they may be connected by one or more paths of length two or more within a's neighborhood, as shown in blue. In fact they may even

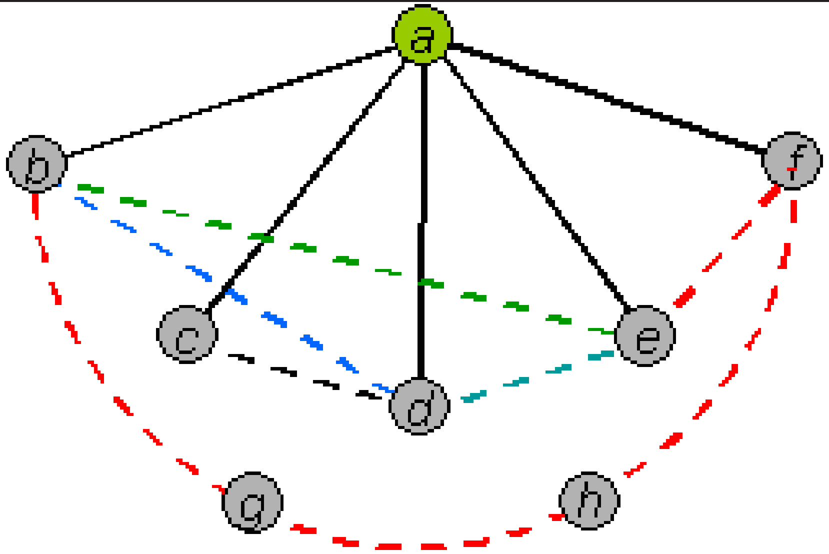

Fig. 5. A phenotype-centered gene neighborhood graph. Vertex connectivity is used to identify genes that are more robust to mutation. Vertex d represents a gene that is in the neighborhood of the phenotype of interest, a, based on genetic correlations in a reference population. Because d is a member of several vertex disjoint paths between b and e, it is more robust to perturbations than other genes in the network.

be connected by one or more paths of length two or more via genes not in a's neighborhood, as shown in red. Because one example of each type is depicted in Figure 5, and because these paths are vertex disjoint, we would in this case say that b and e are 3-connected. The biological interpretation is that three distinct pathways join b and e, thereby providing redundancy, around d. Thus, mutations in d are likely to be accommodated by other alternate pathways. We will use a min-max characterization to score a gene in the neighborhood of a phenotype. To illustrate, let x, y and z denote three such genes with $x \neq y \neq z$. The score of x relative to y and z, denoted $R_{yz}(x)$, is the maximum number of vertex-disjoint paths, exclusive of a, that remain to connect y and z if x is eliminated from the phenotype-centered-gene-neighborhood graph. The robustness score of x, denoted $R(x)$, is the minimum value of $R_{yz}(x)$ taken over all pairs y and z for which x lies on some path between y and z. In figure 5, for example, $R_{be}(d) = 2$, because the elimination of d leaves the two paths $b - e$ and $b - g - h - f - e$. But all paths between c and e must go through d, and so $R_{ce}(d) = R(d) = 0$.

8 Vertex Coverage and Gene-to-Phenotype Networks

Determining the relationship between higher order phenotypes and paracliques and other dense subgraphs in the genetic co-expression network is made readily possible using the reference population. Over 1500 phenotypes have been obtained in mouse recombinant inbred strains. The phenotypes are collected in various subsets of the strain population, and unlike the gene expression correlations, may contain substantial missing in a non-uniform pattern across the data matrix. We have thus used p-value for the genetic correlation as an edge weight threshold. An example is shown in Figure 6, in which a small paraclique of brain transcript abundances containing an inwardly rectifying voltage-gated potassium channel, *Kcnj9*, is heavily involved in behavioral variation and shown to be related to blood alcohol at the return of the righting reflex and preference for sweet solutions.

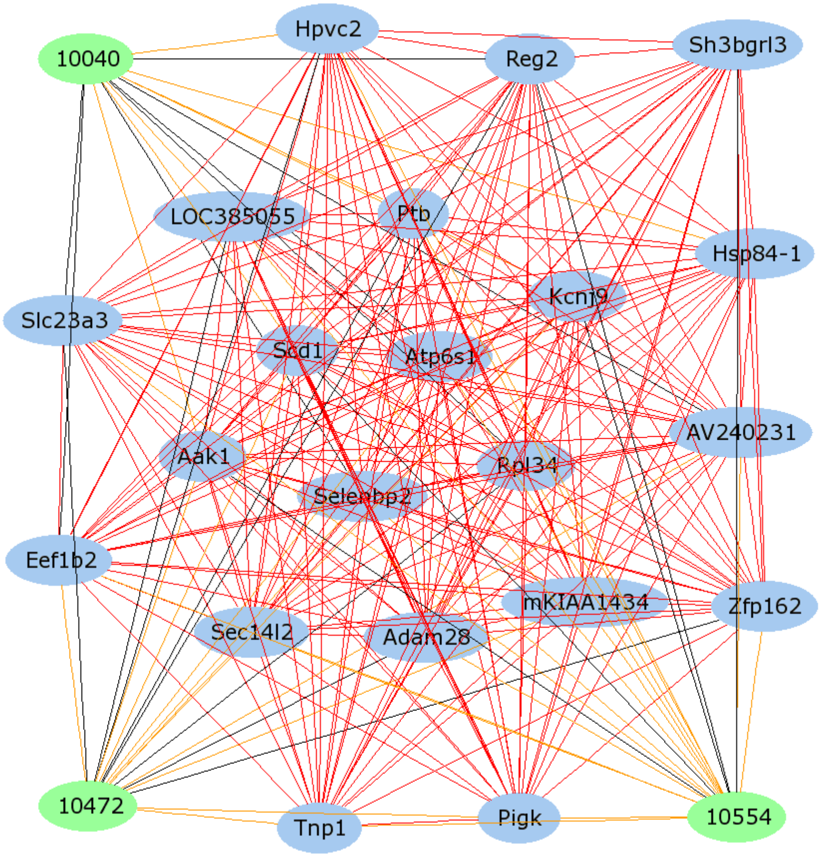

Fig. 6. At an edge weight threshold of $p < .05$, the phenotypes 10472 and 10554, shown in green, cover 100% of the paraclique members. Phenotype 10040 covers 95.42% of the paraclique members. Phenotype vertex labels are as follows: 10040 - BEC at return of Righting Reflex 10472, 10554 - Saccaharin preference basal and after EtOH.

9 Scalability, Data Dimensionality, and Directions for Future Research

An exciting potential of systems genetics is continually to aggregate biological data in reference populations across all levels of biological scale in a highly complex multicellular organism, the laboratory mouse. Already, there are massive numbers of unique genotype vectors in several existing reference populations, and expression assays for thousands of genes in at least ten tissue types. At present, there are approximately 1500 higher order phenotypes available in these populations, and many more in collection. As the reliability and precision of high-throughput proteomics and cell-culture assays improves, the amount of data available in these lines will increase markedly. Additional data will produce multiplicative growth in the number of correlations that are defined. While linear modeling and other parametric approaches have been applied with much success for known pathways, extracting novel information from data of this scale is a phenomenal challenge. Discretizing this complex correlational system and applying advances in graph algorithms have given us an efficient and relatively rapid means for reducing the data dimensionality and extracting networks of genes and phenotypes. The approaches that we have illustrated are highly scalable, and are capable of extracting large sets of related traits from the entire relational system.

In addition to expansive volumes of data, there is a growing complexity to the types of research questions that can be asked. We are presently developing approaches to compare graphs collected in a systems genetic context to reflect differences in time, tissue and treatment effects. Visualization methods and compelling biological validation of novel results are essential to translate these methods and deliver them to the broader audience of biologists who are already successfully harnessing the insight into specific gene-regulatory relations that these public data sets have allowed.

Acknowledgments

This work has been supported in part by the National Science Foundation under grant CCR-0311500, by the Office of Naval Research under grant N00014-01-1-0608, and by the National Institutes of Health under grants 1-P01-DA-015027-01 and 1-R01-MH-074460-01. It has employed resources managed by UT-Battelle for the U.S. Department of Energy under contract DE-AC05-00OR22725. We thank Dr. Robert W. Williams and Dr. Lu Lu of the University of Tennessee Health Science Center for providing data for the analyses we have described here, and Evan G. Williams for rendering Figure 3.

References

1. F. N. Abu-Khzam, R. L. Collins, M. R. Fellows, M. A. Langston, W. H. Suters, and C. T. Symons. Kernelization algorithms for the vertex cover problem: Theory and experiments. In *Proceedings, Workshop on Algorithm Engineering and Experiments*, New Orleans, Louisiana, 2004.
2. F. N. Abu-Khzam, M. A. Langston, P. Shanbhag, and C. T. Symons. Scalable parallel algorithms for FPT problems. *Algorithmica*, 2006, accepted for publication.
3. O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97:10101–10106, 2000.

4. N. E. Baldwin, E. J. Chesler, S. Kirov, M. A. Langston, J. R. Snoddy, R. W. Williams, and B. Zhang. Computational, integrative, and comparative methods for the elucidation of genetic coexpression networks. *Journal of Biomedicine and Biotechnology*, 2:172–180, 2005.
5. M. Bartoli, J. P. Ternaux, C. Forni, P. Portalier, P. Salin, M. Amalric, and A. Monneron. Down-regulation of striatin, a neuronal calmodulin-binding protein, impairs rat locomotor activity. *Journal of Neurobiology*, 40:234–243, 1999.
6. C. Becamel, S. Gavarini, B. Chanrion, G. Alonso, N. Galeotti, A. Dumuis, J. Bockaert, and P. Marin. The serotonin 5-ht2a and 5-ht2c receptors interact with specific sets of pdz proteins. *Journal of Biological Chemistry*, 279:20257–20266, 2004.
7. A. Bellaachia, D. Portnoy, Y. Chen, and A. G. Elkahloun. E-cast: A data mining algorithm for gene expression data. In *Proceedings, Workshop on Data Mining in Bioinformatics*, Edmonton, Alberta, Canada, 2002.
8. A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *Journal of Computational Biology*, pages 54–64, 2000.
9. A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4):281–297, 1999.
10. I. Bomze, M. Budinich, P. Pardalos, and M. Pelillo. The maximum clique problem. In D. Z. Du and P. M. Pardalos, editors, *Handbook of Combinatorial Optimization*, volume 4. Kluwer Academic Publishers, 1999.
11. R. B. Brem and L. Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences*, 102:1572–1577, 2005.
12. R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296:752–755, 2002.
13. K. W. Broman and T. P. Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:641–656, 2002.
14. A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97:12182–12186, 2000.
15. S. Butz, M. Okamoto, and T. C. Sudhof. A tripartite protein complex with the potential to couple synaptic vesicle exocytosis to cell adhesion in brain. *Cell*, 94:773–782, 1998.
16. L. Bystrykh, E. Weersing, B. Dontje, S. Sutton, M. T. Pletcher, T. Wiltshire, A. Su, E. Vellenga, J. Wang, K. F. Manly, L. Lu, E. J. Chesler, R. Alberts, R. C. Jansen, R. W. Williams, M. P. Cooke, and G. d. Haan. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nature Genetics*, 37:225–232, 2005.
17. L. S. Chandran and F. Grandoni. Refined memorisation for vertex cover. In *Proceedings, International Workshop on Parameterized and Exact Computation (IWPEC)*, 2004.
18. E. J. Chesler, L. Lu, S. Shou, Y. Qu, J. Gu, J. Wang, H. C. Hsu, J. D. Mountz, N. E. Baldwin, M. A. Langston, J. B. Hogenesch, D. W. Threadgill, K. F. Manly, and R. W. Williams. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics*, 37:233–242, 2005.
19. E. J. Chesler, L. Lu, J. Wang, R. W. Williams, and K. F. Manly. Webqtl: Rapid exploratory analysis of gene expression and genetic networks for brain and behavior. *Nature Neuroscience*, 7:486–486, 2004.
20. E. J. Chesler, J. Wang, L. Lu, Y. Qu, K. F. Manly, and R. W. Williams. Genetic correlates of gene expression in recombinant inbred strains: a relational model system to explore neurobehavioral phenotypes. *Neuroinformatics*, 1:343–357, 2003.
21. E. J. Chesler and R. W. Williams. Brain gene expression: Genomics and genetics. *International Review of Neurobiology*, 60:59–95, 2004.

22. G. A. Churchill, D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie, J. Beatty, W. D. Beavis, J. K. Belknap, B. Bennett, W. Berrettini, A. Bleich, M. Bogue, K. W. Broman, K. J. Buck, E. Buckler, M. Burmeister, E. J. Chesler, J. M. Cheverud, S. Clapcote, M. N. Cook, R. D. Cox, J. C. Crabbe, W. E. Crusio, A. Darvasi, C. F. Deschepper, R. W. Doerge, C. R. Farber, J. Forejt, D. Gaile, S. J. Garlow, H. Geiger, H. Gershenfeld, T. Gordon, J. Gu, W. Gu, G. d. Haan, N. L. Hayes, C. Heller, H. Himmelbauer, R. Hitzemann, K. Hunter, H. C. Hsu, F. A. Iraqi, B. Ivandic, H. J. Jacob, R. C. Jansen, K. J. Jepsen, D. K. Johnson, T. E. Johnson, G. Kempermann, C. Kendziorski, M. Kotb, R. F. Kooy, B. Llamas, F. Lammert, J. M. Lassalle, P. R. Lowenstein, A. L. L. Lu, K. F. Manly, R. Marcucio, D. Matthews, J. F. Medrano, D. R. Miller, G. Mittleman, B. A. Mock, J. S. Mogil, X. Montagutelli, G. Morahan, D. G. Morris, R. Mott, J. H. Nadeau, H. Nagase, R. S. Nowakowski, B. F. O'Hara, A. V. Osadchuk, G. P. Page, A. Paigen, K. Paigen, A. A. Palmer, H. J. Pan, L. Peltonen-Palotie, J. Peirce, D. Pomp, M. Pravenec, D. R. Prows, Z. Qi, R. H. Reeves, J. Roder, G. D. Rosen, E. E. Schadt, L. C. Schalkwyk, Z. Seltzer, K. Shimomura, S. Shou, M. J. Sillanpaa, L. D. Siracusa, H. W. Snoeck, J. L. Spearow, K. Svenson, L. M. Tarantino, D. Threadgill, L. A. Toth, W. Valdar, F. P. d. Villena, C. Warden, S. Whatley, R. W. Williams, T. Wiltshire, N. Yi, D. Zhang, M. Zhang, and F. Zou. The collaborative cross, a community resource for the genetic analysis of complex traits. *Nature Genetics*, 36:1133–1137, 2004.
23. R. W. Doerge. Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics*, 3:43–52, 2002.
24. R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Springer-Verlag, 1999.
25. U. Feige, D. Peleg, and G. Kortsarz. The dense k-subgraph problem. *Algorithmica*, 29:410–421, 2001.
26. M. Girolami and R. Breitling. Biologically valid linear factor models of gene expression. *Bioinformatics*, 20:3021–3033, 2004.
27. P. Hansen and B. Jaumard. Cluster analysis and mathematical programming. *Mathematical Programming*, 79(1-3):191–215, 1997.
28. E. Hartuv, A. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrachs, and R. Shamir. An algorithm for clustering cDNAs for gene expression analysis. In *Proceedings, RECOMB,* Lyon, France, 1999.
29. L. J. Heyer, S. Kruglyak, and S. Yooseph. Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research*, 9:1106–1115, 1999.
30. N. Hubner, C. A. Wallace, H. Zimdahl, E. Petretto, H. Schulz, F. Maciver, M. Mueller, O. Hummel, J. Monti, V. Zidek, A. Musilova, V. Kren, H. Causton, L. Game, G. Born, S. Schmidt, A. Muller, S. A. Cook, T. W. Kurtz, J. Whittaker, M. Pravenec, and T. J. Aitman. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics*, 37:243–253, 2005.
31. M. A. Langston, L. Lan, X. Peng, N. E. Baldwin, C. T. Symons, B. Zhang, and J. R. Snoddy. A combinatorial approach to the analysis of differential gene expression data: The use of graph algorithms for disease prediction and screening. In J. S. Shoemaker and S. M. Lin, editors, *Methods of Microarray Data Analysis IV*. Springer Verlag, 2005.
32. M. A. Langston, A. D. Perkins, A. M. Saxton, J. A. Scharff, and B. H. Voy. Innovative computational methods for transcriptomic data analysis. In *Proceedings, ACM Symposium on Applied Computing*, Dijon, France, 2006, accepted for publication.
33. J. Li and M. Burmeister. Genetical genomics: Combining genetics with gene expression analysis. *Human Molecular Genetics*, 14:163–169, 2005.
34. K. F. Manly and J. M. Olson. Overview of qtl mapping software and introduction to map manager qt. *Mammalian Genome*, 10:327–334, 1999.
35. J. L. Peirce, L. Lu, J. Gu, L. M. Silver, and R. W. Williams. A new set of bxd recombinant inbred lines from advanced intercross populations in mice. *BMC Genetics*, 5:7, 2004.

36. E. E. Schadt, S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, P. S. Linsley, M. Mao, R. B. Stoughton, and S. H. Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422:297–302, 2003.
37. D. K. Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nature*, 32:502–508, 2002.
38. A. Wagner. Distributed robustness versus redundancy as causes of mutational robustness. *Bioessays*, 27:176–188, 2005.
39. Y. Zhang, F. N. Abu-Khzam, N. E. Baldwin, E. J. Chesler, M. A. Langston, and N. F. Samatova. Genome-scale computational approaches to memory-intensive applications in systems biology. In *Proceedings, Supercomputing*, Seattle, Washington, 2005.